IN PARTNERSHIP WITH:
CNRS

Université de Bordeaux

Activity Report 2013

Project-Team MAGNOME

Models and Algorithms for the Genome

IN COLLABORATION WITH: Laboratoire Bordelais de Recherche en Informatique (LaBRI)

RESEARCH CENTER
Bordeaux - Sud-Ouest

THEME
Computational Biology

Table of contents

Project-Team MAGNOME

Keywords: Computational Biology, Genomics, Knowledge Engineering, Modeling, High Performance Computing

Creation of the Project-Team: 2009 July 01.

1. Members

Research Scientists

David Sherman [Team leader, Inria, Senior Researcher, HdR]
Pascal Durrens [CNRS, Researcher, HdR]

Engineers

Xavier Calcas [CNRS, from Jun 2013]
Florian Lajus [Inria]

PhD Students

Razanne Issa [Exchange Fellowship Syria]
Anna Zhukova [Inria]

Post-Doctoral Fellow

Witold Dyrka [Inria, funded by ANR Mykimum project]

Visiting Scientists

Natalia Golenetskaya [R&D Scientist, until Jun 2013]
Artem Kasianov [PhD student RAS Moscow, from Oct 2013 until Nov 2013]
Vsevolod Makeev [Professor RAS Moscow, until Nov 2013]

Administrative Assistant

Anne-Laure Gautier [Inria]

Other

Joaquin Francisco Fernandez [Inria, PhD student U.Rosario, from Sep 2013 until Nov 2013]

2. Overall Objectives

2.1. Overall Objectives

One of the key challenges in the study of biological systems is understanding how the static information recorded in the genome is interpreted to become dynamic systems of cooperating and competing biomolecules. MAGNOME addresses this challenge through the development of informatic techniques for multi-scale modeling and large-scale comparative genomics:

- logical and object models for knowledge representation
- stochastic hierarchical models for behavior of complex systems, formal methods
- algorithms for sequence analysis, and
- data mining and classification.

We use genome-scale comparisons of eukaryotic organisms to build modular and hierarchical hybrid models of cell behavior that are studied using multi-scale stochastic simulation and formal methods. Our research program builds on our experience in comparative genomics, modeling of protein interaction networks, and formal methods for multi-scale modeling of complex systems.

New high-throughput technologies for DNA sequencing have radically reduced the cost of acquiring genome and transcriptome data, and introduced new strategies for whole genome sequencing. The result has been an increase in data volumes of several orders of magnitude, as well has a greatly increased density of genome sequences within phylogenetically constrained groups of species. MAGNOME develops efficient techniques for dealing with these increased data volumes, and the combinatorial challenges of dense multi-genome comparison.

3. Research Program

3.1. Overview

Fundamental questions in the life sciences can now be addressed at an unprecedented scale through the combination of high-throughput experimental techniques and advanced computational methods from the computer sciences. The new field of *computational biology* or *bioinformatics* has grown around intense collaboration between biologists and computer scientists working towards understanding living organisms as *systems*. One of the key challenges in this study of systems biology is understanding how the static information recorded in the genome is interpreted to become dynamic systems of cooperating and competing biomolecules.

MAGNOME addresses this challenge through the development of informatic techniques for understanding the structure and history of eukaryote genomes: algorithms for genome analysis, data models for knowledge representation, stochastic hierarchical models for behavior of complex systems, and data mining and classification. Our work is in methods and algorithms for:

- **Genome annotation** for complete genomes, performing *syntactic* analyses to identify genes, and *semantic* analyses to map biological meaning to groups of genes [35], [6], [10], [11], [49], [50].
- **Integration of heterogenous data**, to build complete knowledge bases for storing and mining information from various sources, and for unambiguously exchanging this information between knowledge bases [1], [4], [41], [44], [33].
- **Ancestor reconstruction** using optimization techniques, to provide plausible scenarios of the history of genome evolution [11], [8], [45], [54].
- **Classification and logical inference**, to reliably identify similarities between groups of genetic elements, and infer rules through deduction and induction [9], [7], [10].
- **Hierarchical and comparative modeling**, to build mathematical models of the behavior of complex biological systems, in particular through combination, reutilization, and specialization of existing continuous and discrete models [40], [30], [53], [37], [52].

The hundred- to thousand-fold decrease in sequencing costs seen in the past few years presents significant challenges for data management and large-scale data mining. MAGNOME's methods specifically address "scaling out," where resources are added by installing additional computation nodes, rather than by adding more resources to existing hardware. Scaling out adds capacity and redundancy to the resource, and thus fault tolerance, by enforcing data redundancy between nodes, and by reassigning computations to existing nodes as needed.

3.2. Comparative genomics

The central dogma of evolutionary biology postulates that contemporary genomes evolved from a common ancestral genome, but the large scale study of their evolutionary relationships is frustrated by the unavailability of these ancestral organisms that have long disappeared. However, this common inheritance allows us to discover these relationships through *comparison*, to identify those traits that are common and those that are novel inventions since the divergence of different lineages.

We develop efficient methodologies and software for associating biological information with complete genome sequences, in the particular case where several phylogenetically-related eukaryote genomes are studied simultaneously.

The methods designed by MAGNOME for comparative genome annotation, structured genome comparison, and construction of integrated models are applied on a large scale to:

- eukaryotes from the hemiascomycete class of yeasts [49], [50], [6], [10], [2], [11] and to
- prokaryotes from the lactic bacteria used in winemaking [35], [36], [43], [34], [38], [32].

3.3. Comparative modeling

A general goal of systems biology is to acquire a detailed quantitative understanding of the dynamics of living systems. Different formalisms and simulation techniques are currently used to construct numerical representations of biological systems, and a recurring challenge is that hand-tuned, accurate models tend to be so focused in scope that it is difficult to repurpose them. We claim that, instead of modeling individual processes *de novo*, a sustainable effort in building efficient behavioral models must proceed incrementally. *Hierarchical modeling* is one way of combining specific models into networks. Effective use of hierarchical models requires both formal definition of the semantics of such composition, and efficient simulation tools for exploring the large space of complex behaviors. We have combined uses theoretical results from formal methods and practical considerations from modeling applications to define BioRica [27], [40], [53], a framework in which discrete and continuous models can communicate with a clear semantics. Hierarchical models in BioRica can be assembled from existing models, and translated into their execution semantics and then simulated at multiple resolutions through multi-scale stochastic simulation. BioRica models are compiled into a discrete event formalism capable of capturing discrete, continuous, stochastic, non deterministic and timed behaviors in an integrated and non-ambiguous way. Our long-term goal to develop a methodology in which we can **assemble a model** for a species of interest using a library of reusable models and a organism-level “schematic” determined by comparative genomics.

Comparative modeling is also a matter of reconciling experimental data with models [5] [30] and inferring new models through a combination of comparative genomics and successive refinement [46], [47].

4. Application Domains

4.1. Function and history of genomes

Yeasts provide an ideal subject matter for the study of eukaryotic microorganisms. From an experimental standpoint, the yeast *Saccharomyces cerevisiae* is a model organism amenable to laboratory use and very widely exploited, resulting in an astonishing array of experimental results. From a genomic standpoint, yeasts from the hemiascomycete class provide a unique tool for studying eukaryotic genome evolution on a large scale. With their relatively small and compact genomes, yeasts offer a unique opportunity to explore eukaryotic genome evolution by comparative analysis of several species. MAGNOME applies its methods for comparative genomics and knowledge engineering to the yeasts through the ten-year old Génolevures program (GDR 2354 CNRS), devoted to large-scale comparisons of yeast genomes with the aim of addressing basic questions of molecular evolution.

We developed the software tools used by the CNRS’s http://www.genolevures.org/ web site. For example, MAGNOME’s Magus system for simultaneous genome annotation combines semi-supervised classification and rule-based inference in a collaborative web-based system that explicitly uses comparative genomics to simultaneously analyse groups of related genomes.

4.2. Alternative fuels and bioconversion

Oleaginous yeasts are capable of synthesizing lipids from different substrates other than glucose, and current research is attempting to understand this conversions with the goal of optimizing their throughput, production and quality. From a genomic standpoint the objective is to characterize genes involved in the biosynthesis of precursor molecules which will be transformed into fuels, which are thus not derived from petroleum.

MAGNOME's focus is in acquiring genome sequences, predicting genes using models learned from genome comparison and sequencing of cDNA transcripts, and comparative annotation. Our overall goal is to define dynamic models that can be used to predict the behavior of modified strains and thus drive selection and genetic engineering.

4.3. Winemaking and improved strain selection

Yeasts and bacteria are essential for the winemaking process, and selection of strains based both on their efficiency and on the influence on the quality of wine is a subject of significant effort in the Aquitaine region. Unlike the species studied above, yeast and bacterial starters for winemaking cannot be genetically modified. In order to propose improved and more specialized starters, industrial producers use breeding and selection strategies.

Comparative genomics is a powerful tool for strain selection even when genetic engineering must be excluded. Large-scale comparison of the genomes of experimentally characterized strains can be used to identify quantitative trait loci, which can be used as markers in selective breeding strategies. Identifying individual SNPs and predicting their effect can lead to better understanding of the function of genes implicated in improved strain performance, particularly when those genes are naturally mutated or are the result of the transfer of genetic material from other strains. And understanding the combined effect of groups of genes or alleles can lead to insight in the phenomenon of heterosis.

4.4. Knowledge bases for molecular tools

Affinity binders are molecular tools for recognizing protein targets, that play a fundamental in proteomics and clinical diagnostics. Large catalogs of binders from competing technologies (antibodies, DNA/RNA aptamers, artificial scaffolds, etc.) and Europe has set itself the ambitious goal of establishing a comprehensive, characterized and standardized collection of specific binders directed against all individual human proteins, including variant forms and modifications. Despite the central importance of binders, they presently cover only a very small fraction of the proteome, and even though there are many antibodies against some targets (for example, > 900 antibodies against p53), there are none against the vast majority of proteins. Moreover, widely accepted standards for binder characterization are virtually nonexistent. Alongside the technical challenges in producing a comprehensive binder resource are significant logistical challenges, related to the variety of producers and the lack of reliable quality control mechanisms. As part of the ProteomeBinders and Affinomics projects, MAGNOME works to develop knowledge engineering techniques for storing, exploring, and exchanging experimental data used in affinity binder characterization.

5. Software and Platforms

5.1. Magus: Genome exploration and analysis

Participants: David James Sherman [correspondant], Pascal Durrens, Natalia Golenetskaya, Florian Lajus, Xavier Calcas.

The MAGUS genome annotation system integrates genome sequences and sequences features, *in silico* analyses, and views of external data resources into a familiar user interface requiring only a Web navigator. MAGUS implements annotation workflows and enforces curation standards to guarantee consistency and integrity. As a novel feature the system provides a workflow for simultaneous annotation of related genomes through the use of protein families identified by *in silico* analyses; this results in an n-fold increase in curation speed, compared to curation of individual genes. This allows us to maintain standards of high-quality manual annotation while efficiently using the time of volunteer curators. For more information see the MAGUS Gforge web site. [1] MAGUS 1.x is mature software used since 2006 by our collaboration partners. MAGUS 2.0 is developed in an Inria Technology Development Action (ADT) with an open-source license and is being deposited with the APP.

[1] http://magus.gforge.inria.fr

5.2. Pantograph: Inference of metabolic networks

Participants: David James Sherman [correspondant], Pascal Durrens, Nicolás Loira, Anna Zhukova.

Pantograph is a software tool for inferring whole-genome metabolic models for eukaryote cell factories. It is based on metabolic scaffolds, abstract descriptions of reactions and pathways on which inferred reactions are hung are are eventually connected by an interative mapping and specialization process. Scaffold fragments can be repeatedly used to build specialized subnetworks of the complete model. A novel feature of Pantograph is that it uses expert knowledge implicitly encoded in the scaffold's gene associations, and explicitly transfers this knowledge to the new model. Pantograph is available under an open-source license. For more information see the Pantograph Gforge web site. [2].

5.3. MetaModGen: Generalizing Metabolic Models

Participants: Anna Zhukova [correspondant], David James Sherman.

The metabolic model generalization and navigation software allows a human expert to explore a metabolic model in a layered manner. The software creates an on-line semantically zoomable representation of a model submitted by the user in SBML [3] format. The most general view represents the compartments of the model; the next view shows the visualization of generalized versions of reactions and metabolites in each compartment (see section 6.3); and the most detailed view visualizes the initial model with the generalization-based layout (where similar metabolites and reactions are placed next to each other). Zoomable representation is implemented using the Leaflet[4] JavaScript library for mobile-friendly interactive maps. Users can click on reactions and compounds to see the information about their annotations. An example of a zoomable representation of the peroxisome compartment of *Y. lipolytica* is available at http://metamogen.gforge.inria.fr/map.html.

5.4. BioRica: Multi-scale Stochastic Modeling

Participants: David James Sherman [correspondant], Rodrigo Assar Cuevas, Joaquin Fernandez.

BioRica is a high-level modeling framework integrating discrete and continuous multi-scale dynamics within the same semantics field. A model in BioRica node is hierarchically composed of nodes, which may be existing models. Individual nodes can be of two types:

- Discrete nodes are composed of states and transitions described by guarded events. Behavior can be stochastic (defined by the likelihood that an event fires when activated) and timed (defined by the delay between an event's activation and the moment that its transition occurs).
- Continuous nodes are described by ODE systems, potentially a hybrid system whose internal state flows continuously while having discrete jumps.

The system has been implemented as a distributable software package. The BioRica compiler reads a specification for hierarchical model and compiles it into an executable simulator. The modeling language is a stochastic extension to the AltaRica [5] Dataflow language, inspired by work of Antoine Rauzy. Input parsers for SBML 2 version 4 are curently being validated. The compiled code uses the Python runtime environment and can be run stand-alone on most systems. For more information see the BioRica Gforge web site. [6] BioRica was developed as an Inria Technology Development Action (ADT) with an open-source license and is deposited with the APP.

5.5. Génolevures On Line: Comparative Genomics of Yeasts

Participants: Pascal Durrens [correspondant], Natalia Golenetskaya, Tiphaine Martin, David James Sherman.

[2] http://pathtastic.gforge.inria.fr
[3] http://sbml.org
[4] http://leafletjs.com
[5] http://altarica.labri.fr
[6] http://biorica.gforge.inria.fr

The Génolevures online database provides tools and data for exploring the annotated genome sequences of more than 20 genomes, determined and manually annotated by the Génolevures Consortium to facilitate comparative genomic studies of hemiascomycetous yeasts. Data are presented with a focus on relations between genes and genomes: conservation of genes and gene families, speciation, chromosomal reorganization and synteny. The Génolevures site includes a private collaboration area for specific studies by members of its international community. The contents of the knowledge base are expanded and maintained by the CNRS through GDR 2354 Génolevures, and full data may be downloaded from the site. Génolevures online uses our open-source MAGUS system for genome navigation, with project-specific extensions developed by David Sherman, Pascal Durrens, and Tiphaine Martin; these extensions are not made available due to incertainty about intellectual property rights. For more information see the Génolevures web site. [7]

5.6. Inria Bioscience Resources

Participants: Olivier Collin [correspondant], Frédéric Cazals, Mireille Régnier, Marie-France Sagot, Hélène Touzet, Hidde De jong, David James Sherman, Marie-Dominique Devignes, Dominique Lavenier.

Inria Bioscience Resources is a portal designed to improve the visibility of bioinformatics tools and resources developed by Inria teams. This portal will help the community of biologists and bioinformatians understand the variety of bioinformatics projects in Inria, test the different applications, and contact project-teams. Eight project-teams participate in the development of this portal. Inria Bioscience Resources is developed in an Inria Technology Development Action (ADT).

6. New Results

6.1. Adopting new computing paradigms

Participants: David James Sherman [correspondant], Pascal Durrens, Natalia Golenetskaya, Florian Lajus, Xavier Calcas.

Analyses in comparative genomics are characteristically forms of datamining in high-dimension sets of relations between genes and gene products. For every linear increase in genomic data, these relations can grow at worst geometrically.

Natalia Golenetskaya's thesis[12] developed an integrated architecture that we call *Tsvetok*, which combines a novel NoSQL storage schema, domain-specific Map-Reduce algorithms, and existing resources to efficiently handle the fundamentally data-parallel analyses encountered in comparative genomics [48], [42], [51]. Tsvetok components are deployed in MAGNOME's private cloud and have been extensively tested using data and use cases derived from log analyses of the Génolevures web resource. We designed Map-Reduce solutions for the principal whole-genome analyses used by MAGNOME for comparative genomics, in particular new distributed algorithms for systematic identification of gene fusion and fission events in eukaryote genomes, and large-scale consensus clustering for protein families. These examples illustrate two strategies that can be used to scale algorithms in a Map-Reduce setting[12].

1. Converting classical graph-based algorithms with message propagation: instead of traversing a graph, which would incur high latency, information is sent forward in waves, and synchronized later. Some of the intermediate computations may be redundant, but overall running time is minimized.
2. Iterative sampling strategies, which run the standard algorithm on carefully chosen subsets, and later compute a consensus of the intermediate results. The iterations may take some time to converge, but the individual instances can be run within one machine.

Florian Lajus extended the Magus software platform to use the NoSQL storage components in Tsvetok, and has validated it on a large collection of fungal genomes. Xavier Calcas is currently integrating the Galaxy platform [8] with Magus.

[7] http://www.genolevures.org/

[8] http://usegalaxy.org

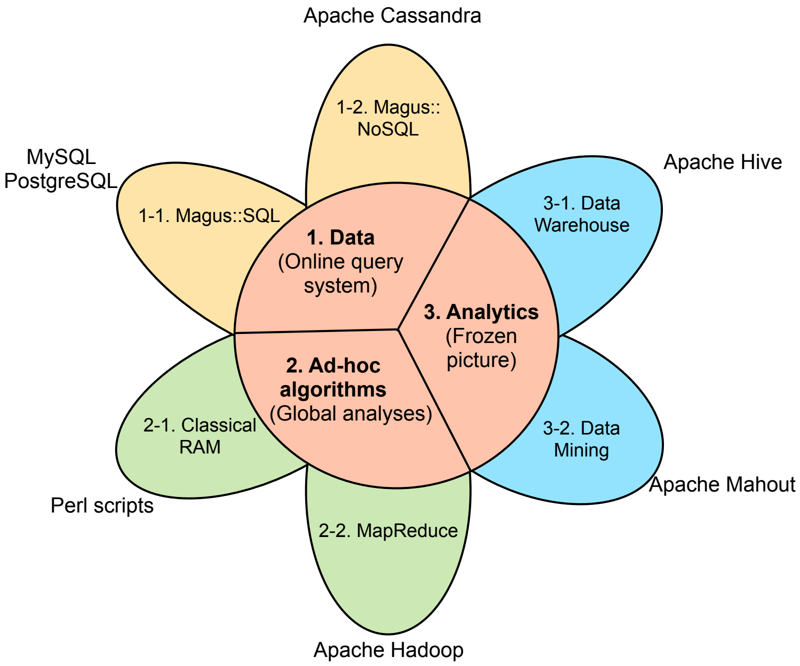

Figure 1. General architecture of Tsvetok, showing the role of NoSQL (Apache Cassandra) and Map-Reduce (Apache Hadoop) paradigms

6.2. Improving inference of metabolic models

Participants: David James Sherman [correspondant], Pascal Durrens, Razanne Issa, Anna Zhukova.

The Pantograph approach uses an annotated “scaffold” (reference) model and a collection of complementary predictions of homology between scaffold genes and target genes. The basis of the method is a weighing of the homology evidence to decide whether a reaction that is present in the scaffold ought be be present in the target.

We have improved on the method in two ways. First, we model the implicit knowledge represented in the boolean formula of each gene association, to derive hypotheses about the explicit role of individual genes; for example, a gene association $(S_1 \wedge S_2) \vee (S_1 \wedge S_3)$ may implicitly represent an enzyme complex formed from two subunits, the first encoded by gene S_1, and the second encoded by two paralogous genes S_2 and S_3 (figure 2). By using these hypotheses to rewrite gene associations, we improve the decision of whether a reaction is present in the target or not.

Second, we have adopted an abductive strategy for inferring reactions. In this strategy we consider that it is the reactions that explain the genes observed in the target genome. In the corresponding abductive logic program, the observations are the genes in the target, the integrity constraints are the rules that rewrite gene associations, and the hypotheses to be abduced are the reactions in the model. The scaffold model is compiled into a set of facts and predicates that express the reactions, their gene associations, and the integrity constraint rules; the abducibles generate assertions that specific reactions are in the target model. Combined with the facts of the genes observed in the target, this program generates, through abduction, the set of target reactions that explain the greatest number of genes.

The advantage of this approach is that it can invent, through specialization, reactions that are not present *per se* in the scaffold model.

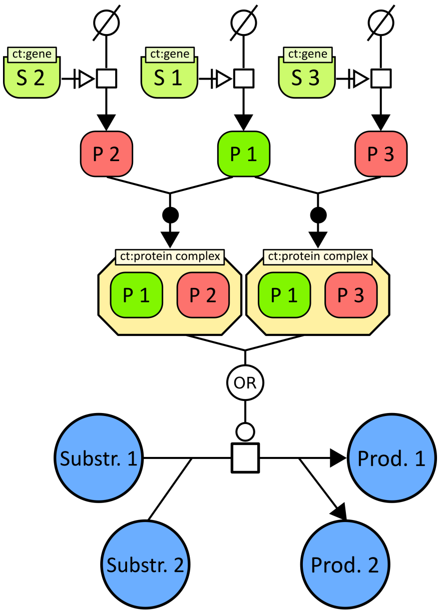

Figure 2. An explicit model that is one possible explanation of the gene association $(S_1 \wedge S_2) \vee (S_1 \wedge S_3)$

6.3. Knowledge-based generalization of metabolic models

Participants: David James Sherman [correspondant], Pascal Durrens, Razanne Issa, Anna Zhukova.

There is an inherent tension between detail and understandability in large metabolic networks: detailed description of individual reactions is needed for simulation, but high-level views of reactions are needed for describing pathways in human terms. We defined knowledge-based methods that factor similar reactions into “generic” reactions in order to visualize a whole pathway or compartment, while maintaining the underlying model so that the user can later “drill down” to the specific reactions if need be[22], [23], [26] This method is available as a Python library from http://metamogen.gforge.inria.fr/.

Figures 3 and 4 illustrate model generation for *Yarrowia lypolitica* fatty acid oxidation in the peroxisome. Molecular species are represented as circular nodes, and the reactions as square ones, connected by edges to their reactants and products. Ubiquitous species (e.g. *oxygen*, *water*, *ATP*) are of smaller size and colored gray. Non-ubiquitous species are divided into fifteen equivalence classes and colored accordingly (red/blue for trivial species/reaction equivalence classes, different colors for non-trivial equivalence classes). The size of the model does not allow for readability of the species labels, thus we do not show them (figure 3).

The generalization algorithm identifies equivalent molecular species using an ontology, and groups together reactions that operate on the same abstract species. It finds the greatest generalization the preserves stoichiometry. The generalized model represents quotient species and reactions. For example, the violet *unsaturated FA-CoA* node is a quotient of *hexadec-2-enoyl-CoA*, *oleoyl-CoA*, *tetradecenoyl-CoA*, *trans-dec-2-enoyl-CoA*, *trans-dodec-2-enoyl-CoA*, *trans-hexacos-2-enoyl-CoA*, *trans-octadec-2-enoyl-CoA*, and *trans-tetradec-2-enoyl-CoA* (colored violet in figure 3). In a similar manner, the light-green *acCoA oxidase* quotient reaction, that converts *fatty acyl-CoA* (yellow) into *unsaturated FA-CoA* (violet), generalizes six corresponding light-green reactions of the initial model (figure 3).

The generalized model describes *β-oxidation* in a more generic way: as a transformation of *fatty acyl-CoA* (yellow) into *unsaturated FA-CoA* (violet), then into *hydroxy FA-CoA* (dark green), *3-oxo FA-CoA* (magenta), and back to *fatty acyl-CoA* (with a shorter carbon chain); while the specific model describes the same process in more details, specifying those reactions for each of the *fatty acyl-CoA* species presented in the organisms' cell (e.g. *decanoyl-CoA*, *dodecanoyl-CoA*, etc.). That is why the *$beta$-oxidation* chain of the reactions in the initial model, transforming step-by-step the *fatty-acyl-CoA* with the longest carbon chain into the one with the shortest chain, in the generalized model appears as a cycle (generalizing all the *fatty-acyl-CoA*s into one species, regardless the chain-length).

The specific model is appropriate for simulation, because it contains all of the precise reactions. The generalized model is suited for a human, because it reveals the main properties of the model and masks distracting details. For example, the generalized model highlights the fact that there is a particularity concerning *C24:0-CoA (stearoyl-CoA)* (red, inside the cycle): there exists a "shortcut" reaction (blue, inside the cycle), producing it directly from another *fatty acyl-CoA* (yellow), avoiding the usual four-reaction beta-oxidation chain, used for other *fatty acyl-CoA*s. This shortcut is not obvious in the specific model, because it is hidden among a plethora of similar-looking reactions.

6.4. Characterization of STAND protein families

Participants: David James Sherman, Pascal Durrens, Witold Dyrka [correspondant].

In collaboration with Sven Saupe and Mathieu Paoletti from IBGC Bordeaux (ANR Mykimun), we worked on characterization of the STAND protein family in the fungal phylum. We established an *in silico* screen based on state-of-the-art bioinformatic tools, which – starting from experimentally studied sequences from *Podospora anserina* – allowed us to determine the first systematic picture of fungal STAND protein repertoire (ms. in preparation). Most notably, we found evidence of extensive modularity of domain associations, and signs of concerted evolution within the recognition domain. Both results support the hypothesis that fungal STAND proteins, originally described in the context of vegetative incompatibility, are involved in a general fungal immune system. In addition, we investigated improved protein domain representations and elaborated a grammatical modelling method [15], which will be used to elucidate mechanisms of formation and operation of the STAND proteins.

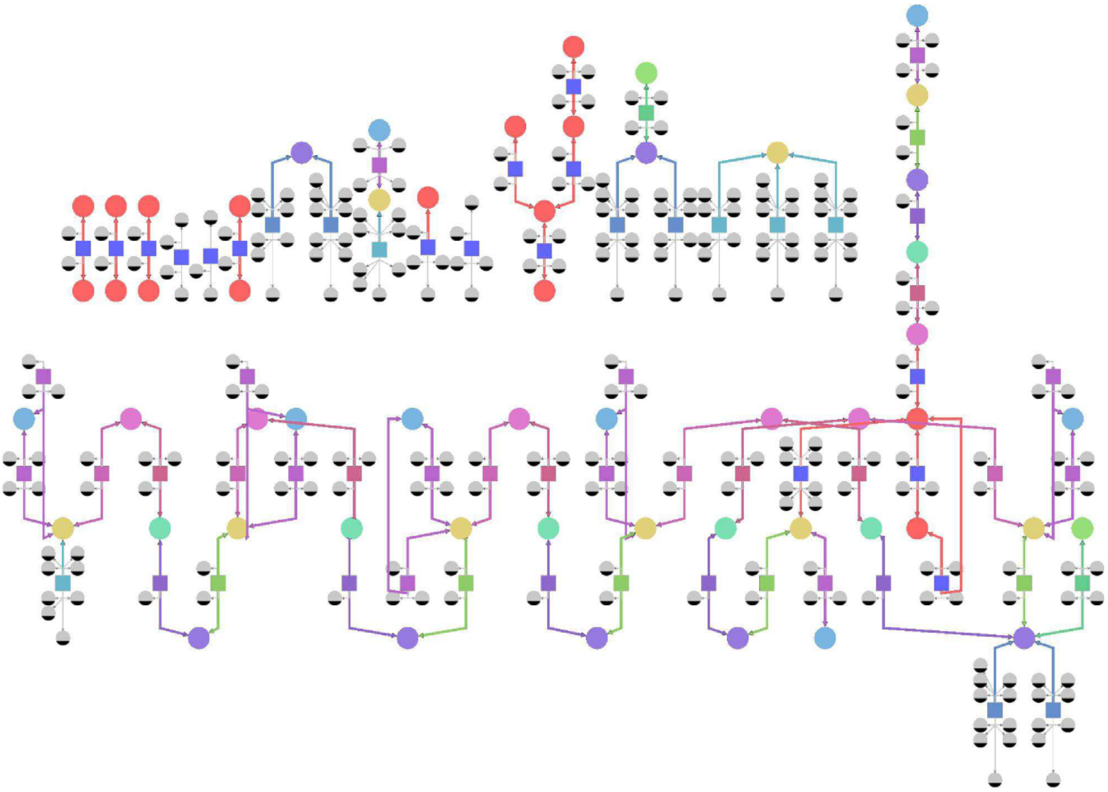

Figure 3. Yarrowia lypolitica fatty acid oxidation model before generalization. Reactions of the specific model are divided into fifteen equivalence classes, represented by different colours

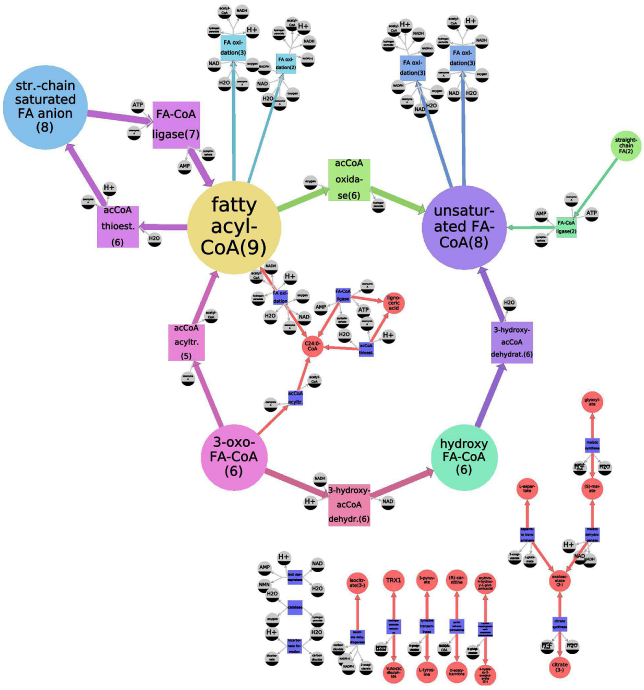

Figure 4. Generalization of the Yarrowia lypolitica fatty acid oxidation model, described as a transformation of fatty acyl-CoA (yellow) into unsaturated FA-CoA (violet), then into hydroxy FA-CoA (dark green), 3-oxo FA-CoA (magenta), and back to fatty acyl-CoA (with a shorter carbon chain)

6.5. Avoiding stiffness in BioRica

Participants: David James Sherman [correspondant], Joaquin Fernandez.

We previously formalized two strategies for integrating discrete control with continuous models, coefficient switches that control the parameters of the continuous model, and strong switches that choose different models [29], [27]. While these strategies have proved useful for modeling hybrid systems in biotechnology [31] and medicine [28], the resulting system model can be inefficient when the different subsystems evolve at very different time scales. In order to improve the efficiency of the resulting simulations, we investigated the use of Kofman's Quantized State Systems (QSS), and demonstrated that the QSS approach can be adapted to BioRica [13]. On the strength of this demonstration, we invited Joaquin Fernandez from Kofman's lab to Magnome. Joaquin had previously implemented an efficient library for QSS simulation, and during his stay succeeded in adapting it to our hybrid modeling framework. In his approach, SBML models with events are compiled into a hybrid model, using a variant of Modelica for surface syntax and using the QSS library for efficient simulation.

6.6. Applications in biotechnology and health

Participants: David James Sherman, Pascal Durrens [correspondant], Florian Lajus, Xavier Calcas.

Using MAGNOME's Magus system and YAGA software, we have successfully realized a full annotation and analysis of several groups of related genomes:

- Seven new genomes, provided to the Génolevures Consortium by the CEA–Génoscope (Évry), including two distant genomes from the *Saccharomycetales* were annotated using previously published Génolevures genomes.
- Twelve wine starter yeasts linked to fermentation efficiency.
- Five pathogenic (to human) and non pathogenic Nakaseomycetes.
- Two oleaginous strains with applications in biofuels.

Winemaking yeasts. In collaboration with partners in the ISVV, Bordeaux, we have assembled and analyzed 12 wine starter yeasts, with the goal of understanding genetic determinants of performance in wine fermentation. Analysis included identification of strain-specific gains and losses of genes linked both to niche specificity and to performance in industrial applications (article in prep.). A further combined analysis with 50 natural and industrial strains showed a pattern of introgression concentrated in industrial strains (article in prep.).

Oleaginous yeasts. In collaboration with Prof Jean-Marc Nicaud's lab at the INRA Grignon, we developed the first functional genome-scale metabolic model of *Yarrowia lipolytica*, an oleaginous yeast studied experimentally for its role as a food contaminant and its use in bioremediation and cell factory applications.

Using MAGNOME's Pantograph method (see section 5.2) we produced an accurate functional model for *Y. lipolytica*, MODEL1111190000 in BioModels [9], that has been qualitatively validated against gene knockouts. This model has been enriched by Anna Zhukova with ontology terms from ChEBI and GO.

Pathogenic yeasts. A further group of five species, comprised of pathogenic and nonpathogenic species, was analyzed with the goal of identifying virulence determinants [39]. By choosing species that are highly related but which differ in the particular traits that are targeted, in this case pathogenicity, we are able to focus of the few hundred genes related to the trait [16]. The approximately 40,000 new genes from these studies were classified into existing Génolevures families as well as branch-specific families.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

MAGNOME and the company BioLaffort are contracted to develop analyses and tools for rationalizing wine starter strain selection using genomics.

[9] http://biomodels.net/

7.2. Bilateral Grants with Industry

The “SAGESS” project, below, section 8.1.1, has been partially funded by a grant to BioLaffort from the Region.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Aquitaine Region “SAGESS” comparative genomics for wine starters.

This project is a collaboration between the company BioLaffort, specialized in the selection of industrial yeasts with distinct technological abilities, with the ISVV and MAGNOME. The goal is to use genome analysis to identify molecular markers responsible for different physiological capabilities, as a tool for selecting yeasts and bacteria for wine fermentation through efficient hybridization and selection strategies. This collaboration has obtained the INNOVIN label.

8.2. National Initiatives

8.2.1. ANR MYKIMUN.

Signal Transduction Associated with Numerous Domains (STAND) proteins play a central role in vegetative incompatibility (VI) in fungi. STAND proteins act as molecular switches, changing from closed inactive conformation to open active conformation upon binding of the proper ligand. Mykimun, coordinated by Mathieu Paoletti of the IBGC (Bordeaux), studies the postulated involvement of STAND proteins in heterospecific non self recognition (innate immune response).

In MYKIMUN we extend the notion of fungal immune receptors and immune reaction beyond the *P. anserina* NWD gene family. We develop *in silico* machine learning tools to identify new potential PRRs based on the expected characteristics of such genes, in *P. anserina* and beyond in additional sequenced fungal genomes. This should contribute to extend concept of a fungal immune system to the whole fungal branch of the eukaryote phylogenetic tree.

8.3. European Initiatives

8.3.1. FP7 Projects

A major objective of the “post-genome” era is to detect, quantify and characterise all relevant human proteins in tissues and fluids in health and disease. This effort requires a comprehensive, characterised and standardised collection of specific ligand binding reagents, including antibodies, the most widely used such reagents, as well as novel protein scaffolds and nucleic acid aptamers. Currently there is no pan-European platform to coordinate systematic development, resource management and quality control for these important reagents.

MAGNOME is an associate partner of the FP7 “Affinity Proteome” project coordinated by Prof. Mike Taussig of the Babraham Institute and Cambridge University. Within the consortium, we participate in defining community for data representation and exchange, and evaluate knowledge engineering tools for affinity proteomics data.

8.3.2. Collaborations with Major European Organizations

Prof. Mike Taussig: Babraham Institute & Cambridge University
Knowledge engineering for Affinity Proteomics
Henning Hermjakob: European Bioinformatics Institute
Standards and databases for molecular interactions

8.4. International Initiatives

8.4.1. Inria Associate Teams

MAGNOME participates in the CARNAGE associated team, coordinated by AMIB, with the Russian Academy of Sciences.

8.4.2. Inria International Partners

8.4.2.1. Declared Inria International Partners

AMAVI

Program: Inria International Partner

Title: Combinatorics and Algorithms for the Genomic sequences

Inria principal investigators: David Sherman

International Partner (Institution - Laboratory - Researcher):

> Vavilov Institute of General Genetics (Russia (Russian Federation)) - Department of Computational Biology - Vsevolod Makeev

Duration: 2010 - present

VIGG and AMIB teams has a more than 12 years long collaboration on sequence analysis. The two groups aim at identifying DNA motifs for a functional annotation, with a special focus on conserved regulatory regions. In the current 3-years project CARNAGE, our collaboration, that includes Inrai-team MAGNOME, is oriented towards new trends that arise from Next Generation Sequencing data. Combinatorial issues in genome assembly are addressed. RNA structure and interactions are also studied.

The toolkit is pattern matching algorithms and analytic combinatorics, leading to common software.

8.4.2.2. Informal International Partners

MAGNOME collaborates with Rodrigo Assar of the Universidad Andrès Bello, and Nicolás Loira and Alessandro Maass of the Center for Genomic Regulation, in Santiago de Chile (Chile).

8.4.3. Participation in other International Programs

MAGNOME and the VIGG of the Russian Academy of Sciences (RAS) in Moscow are partners in a project funded by the CNRS and the RAS entitle “Séquençage profond de organismes biotechnologiques : des régulons à l’adaptation ”.

8.5. International Research Visitors

8.5.1. Visits of International Scientists

Vsevolod MAKEEV November 8-22 2013

Artëm KASIANOV November 8-22 2013

8.5.1.1. Internships

Joaquin FERNANDEZ September-November 2013

9. Dissemination

9.1. Scientific Animation

Pascal Durrens is :

leader of the “Comparative Genomics” theme and member of the Scientific Council of the LaBRI UMR 5800/CNRS.

responsible for scientific diffusion for the Génolevures Consortium.

member of the editorial board of the journal ISRN Computational Biology, and was reviewer for the journal BMC Genomics

expert in Genomics for the Fonds de la Recherche Scientifique-FNRS (FRS-FNRS), Belgium

David Sherman is :

president of the Commission de Jeunes Chercheurs, Inria Bordeaux Sud-Ouest

member for Bordeaux Sud-Ouest of Inria's Young Scientists Mission

member of the editorial board of the journal Computational and Mathematical Methods in Medicine

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Licence : Anna Zhukova, J1MI2013 : Algorithmes et Programmes TD/TP, 30h, L2, Université Bordeaux, France

9.2.2. Supervision

PhD in progress: Anna Zhukova, "Knowledge engineering for biological networks," 2011–, Sherman

PhD in progress: Razanne Issa, "Analyse symbolique de données génomiques," 2010–, Sherman

9.2.3. Juries

David Sherman was a member of the juries of:

Natalia GOLENETSKAYA, "Addressing scaling challenges in comparative genomics," U. Bordeaux, 2013-09-09

Boyang JI, "Comparative and Functional Genome Analysis of Magnetotactic Bacteria," U. Aix-Marseille, 2013-10-23

Andres ARAVENA, "Probabilistic and constraint based modelling to determine regulation events from heterogeneous biological data," U. Rennes, 2013-12-13

9.3. Popularization

Magnome participated in « UniThé ou Café » in the Inria Bordeaux – Sud-Ouest research center.

Anna Zhukova animated one of the Inria workshops at the 2013 "Fête de la Science"

David Sherman is a member of the Inria Bordeaux – Sud-Ouest's "Scientific Culture" committee, which organizes and proposes various scientific popularization actions.

10. Bibliography

Major publications by the team in recent years

[1] R. Barriot, D. J. Sherman, I. Dutour. *How to decide which are the most pertinent overly-represented features during gene set enrichment analysis*, in "BMC Bioinformatics", 2007, vol. 8 [*DOI :* 10.1186/1471-2105-8-332], http://hal.inria.fr/inria-00202721/en/

[2] G. Blandin, P. Durrens, F. Tekaia, M. Aigle, M. Bolotin-Fukuhara, E. Bon, S. Casaregola, J. de Montigny, C. Gaillardin, A. Lépingle, B. Llorente, A. Malpertuy, C. Neuvéglise, O. Ozier-Kalogeropoulos, A. Perrin, S. Potier, J.-L. Souciet, E. Talla, C. Toffano-Nioche, M. Wésolowski-Louvel, C. Marck, B. Dujon. *Genomic Exploration of the Hemiascomycetous Yeasts: 4. The genome of Saccharomyces cerevisiae revisited*, in "FEBS Letters", December 2000, vol. 487, n^{o} 1, pp. 31-36

[3] E. BON, A. DELAHERCHE, E. BILHÈRE, A. DE DARUVAR, A. LONVAUD-FUNEL, C. LE MARREC. *Oenococcus oeni genome plasticity is associated with fitness*, in "Applied and Environmental Microbiology", 2009, vol. 75, n^{o} 7, pp. 2079-90, http://hal.inria.fr/inria-00392015/en/

[4] J. BOURBEILLON, S. ORCHARD, I. BENHAR, C. BORREBAECK, A. DE DARUVAR, S. DÜBEL, R. FRANK, F. GIBSON, D. GLORIAM, N. HASLAM, T. HILTKER, I. HUMPHREY-SMITH, M. HUST, D. JUNCKER, M. KOEGL, Z. KONTHUR, B. KORN, S. KROBITSCH, S. MUYLDERMANS, P.-A. NYGREN, S. PALCY, B. POLIC, H. RODRIGUEZ, A. SAWYER, M. SCHLAPSHY, M. SNYDER, O. STOEVESANDT, M. J. TAUSSIG, M. TEMPLIN, M. UHLEN, S. VAN DER MAAREL, C. WINGREN, H. HERMJAKOB, D. J. SHERMAN. *Minimum information about a protein affinity reagent (MIAPAR)*, in "Nature Biotechnology", 07 2010, vol. 28, n^{o} 7, pp. 650-3 [*DOI :* 10.1038/NBT0710-650], http://hal.inria.fr/inria-00544750/en

[5] A. B. CANELAS, N. HARRISON, A. FAZIO, J. ZHANG, J.-P. PITKÄNEN, J. VAN DEN BRINK, B. M. BAKKER, L. BOGNER, J. BOUWMAN, J. I. CASTRILLO, A. CANKORUR, P. CHUMNANPUEN, P. DARAN-LAPUJADE, D. DIKICIOGLU, K. VAN EUNEN, J. C. EWALD, J. J. HEIJNEN, B. KIRDAR, I. MATTILA, F. I. C. MENSONIDES, A. NIEBEL, M. PENTTILÄ, J. T. PRONK, M. REUSS, L. SALUSJÄRVI, U. SAUER, D. J. SHERMAN, M. SIEMANN-HERZBERG, H. WESTERHOFF, J. DE WINDE, D. PETRANOVIC, S. G. OLIVER, C. T. WORKMAN, N. ZAMBONI, J. NIELSEN. *Integrated multilaboratory systems biology reveals differences in protein metabolism between two reference yeast strains*, in "Nature Communications", 12 2010, vol. 1, n^{o} 9, 145 p. [*DOI :* 10.1038/NCOMMS1150], http://hal.inria.fr/inria-00562005/en/

[6] B. DUJON, D. J. SHERMAN, G. FISCHER, P. DURRENS, S. CASAREGOLA, I. LAFONTAINE, J. DE MONTIGNY, C. MARCK, C. NEUVÉGLISE, E. TALLA, N. GOFFARD, L. FRANGEUL, M. AIGLE, V. ANTHOUARD, A. BABOUR, V. BARBE, S. BARNAY, S. BLANCHIN, J.-M. BECKERICH, E. BEYNE, C. BLEYKASTEN, A. BOISRAMÉ, J. BOYER, L. CATTOLICO, F. CONFANIOLERI, A. DE DARUVAR, L. DESPONS, E. FABRE, C. FAIRHEAD, H. FERRY-DUMAZET, A. GROPPI, F. HANTRAYE, C. HENNEQUIN, N. JAUNIAUX, P. JOYET, R. KACHOURI-LAFOND, A. KERREST, R. KOSZUL, M. LEMAIRE, I. LESUR, L. MA, H. MULLER, J.-M. NICAUD, M. NIKOLSKI, S. OZTAS, O. OZIER-KALOGEROPOULOS, S. PELLENZ, S. POTIER, G.-F. RICHARD, M.-L. STRAUB, A. SULEAU, D. SWENNEN, F. TEKAIA, M. WÉSOLOWSKI-LOUVEL, E. WESTHOF, B. WIRTH, M. ZENIOU-MEYER, I. ZIVANOVIC, M. BOLOTIN-FUKUHARA, A. THIERRY, C. BOUCHIER, B. CAUDRON, C. SCARPELLI, C. GAILLARDIN, J. WEISSENBACH, P. WINCKER, J.-L. SOUCIET. *Genome evolution in yeasts*, in "Nature", 07 2004, vol. 430, n^{o} 6995, pp. 35-44 [*DOI :* 10.1038/NATURE02579], http://hal.archives-ouvertes.fr/hal-00104411/en/

[7] P. DURRENS, M. NIKOLSKI, D. J. SHERMAN. *Fusion and fission of genes define a metric between fungal genomes*, in "PLoS Computational Biology", 10 2008, vol. 4 [*DOI :* 10.1371/JOURNAL.PCBI.1000200], http://hal.inria.fr/inria-00341569/en/

[8] A. GOËFFON, M. NIKOLSKI, D. J. SHERMAN. *An Efficient Probabilistic Population-Based Descent for the Median Genome Problem*, in "Proceedings of the 10th annual ACM SIGEVO conference on Genetic and evolutionary computation (GECCO 2008)", Atlanta United States, ACM, 2008, pp. 315-322, http://hal.archives-ouvertes.fr/hal-00341672/en/

[9] M. NIKOLSKI, D. J. SHERMAN. *Family relationships: should consensus reign?- consensus clustering for protein families*, in "Bioinformatics", 2007, vol. 23 [*DOI :* 10.1093/BIOINFORMATICS/BTL314], http://hal.inria.fr/inria-00202434/en/

[10] D. J. SHERMAN, T. MARTIN, M. NIKOLSKI, C. CAYLA, J.-L. SOUCIET, P. DURRENS. *Genolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes*, in "Nucleic Acids Research (NAR)", 2009, D p. [*DOI :* 10.1093/NAR/GKN859], http://hal.inria.fr/inria-00341578/en/

[11] J.-L. Souciet, B. Dujon, C. Gaillardin, M. Johnston, P. Baret, P. Cliften, D. J. Sherman, J. Weissenbach, E. Westhof, P. Wincker, C. Jubin, J. Poulain, V. Barbe, B. Ségurens, F. Artiguenave, V. Anthouard, B. Vacherie, M.-E. Val, R. S. Fulton, P. Minx, R. Wilson, P. Durrens, G. Jean, C. Marck, T. Martin, M. Nikolski, T. Rolland, M.-L. Seret, S. Casaregola, L. Despons, C. Fairhead, G. Fischer, I. Lafontaine, V. Leh Louis, M. Lemaire, J. De Montigny, C. Neuvéglise, A. Thierry, I. Blanc-Lenfle, C. Bleykasten, J. Diffels, E. Fritsch, L. Frangeul, A. Goëffon, N. Jauniaux, R. Kachouri-Lafond, C. Payen, S. Potier, L. Pribylova, C. Ozanne, G.-F. Richard, C. Sacerdot, M.-L. Straub, E. Talla. *Comparative genomics of protoploid Saccharomycetaceae*, in "Genome Research", 2009, vol. 19, pp. 1696-1709, http://hal.inria.fr/inria-00407511/en/

Publications of the year

Doctoral Dissertations and Habilitation Theses

[12] N. Golenetskaya. , *Adressing scaling challenges in comparative genomics*, Université Sciences et Technologies - Bordeaux I, September 2013, http://hal.inria.fr/tel-00865840

Articles in International Peer-Reviewed Journals

[13] R. Assar, D. J. Sherman. *Implementing biological hybrid systems: Allowing composition and avoiding stiffness*, in "Applied Mathematics and Computation", August 2013, http://hal.inria.fr/hal-00853997

[14] W. Dyrka, M. M. Bartuzel, M. Kotulska. *Optimization of 3D Poisson-Nernst-Planck model for fast evaluation of diverse protein channels*, in "Proteins: Structure, Function, and Bioinformatics", August 2013 [*DOI :* 10.1002/PROT.24326], http://hal.inria.fr/hal-00857213

[15] W. Dyrka, J.-C. Nebel, M. Kotulska. *Probabilistic grammatical model for helix-helix contact site classification*, in "Algorithms for Molecular Biology", 2013, vol. 8, 31 p. [*DOI :* 10.1186/1748-7188-8-31], http://hal.inria.fr/hal-00923291

[16] T. Gabaldón, T. Martin, M. Marcet-Houben, P. Durrens, M. Bolotin-Fukuhara, O. Lespinet, S. Arnaise, S. Boisnard, G. Aguileta, R. Atanasova, C. Bouchier, A. Couloux, S. Creno, J. Almeida Cruz, H. Devillers, A. Enache-Angoulvant, J. Guitard, L. Jaouen, L. Ma, C. Marck, C. Neuvéglise, E. Pelletier, A. Pinard, J. Poulain, J. Recoquillay, E. Westhof, P. Wincker, B. Dujon, C. Hennequin, C. Fairhead. *Comparative genomics of emerging pathogens in the Candida glabrata clade*, in "BMC Genomics", September 2013, vol. 14, n^{o} 1, 623 p. [*DOI :* 10.1186/1471-2164-14-623], http://hal.inria.fr/inserm-00871184

[17] A. Romano, H. Trip, H. Campbell-Sills, O. Bouchez, D. D. Sherman, J. S. Lolkema, P. M. Lucas. *Genome Sequence of Lactobacillus saerimneri 30a (Formerly Lactobacillus sp. Strain 30a), a Reference Lactic Acid Bacterium Strain Producing Biogenic Amines*, in "Genome Announcements", January 2013, vol. 1, n^{o} 1, 12 p. [*DOI :* 10.1128/GENOMEA.00097-12], http://hal.inria.fr/hal-00863284

[18] A. Sarkar, M. Nikolski, P. Durrens. *The family based variability in protein family expansion*, in "International Journal of Bioinformatics Research and Applications", 2013, vol. 9, n^{o} 2, pp. 121-33 [*DOI :* 10.1504/IJBRA.2013.052473], http://hal.inria.fr/hal-00857374

[19] A. Zhukova, D. J. Sherman. *Knowledge-based generalization of metabolic models*, in "Journal of Computational Biology", 2014, http://hal.inria.fr/hal-00925881

[20] A. ZHUKOVA, D. J. SHERMAN. *Knowledge-based generalization of metabolic networks: a practical study*, in "Journal of Bioinformatics and Computational Biology", 2014, http://hal.inria.fr/hal-00906911

Invited Conferences

[21] D. J. SHERMAN. *Taming the complexity of 'n-ary' relations in comparative genomics*, in "9th International Conference on Genome Biology and Bioinformatics", Atlanta, Georgia, United States, M. BORODOVSKY (editor), Georgia Tech and Emory University, November 2013, http://hal.inria.fr/hal-00938262

International Conferences with Proceedings

[22] A. ZHUKOVA, D. J. SHERMAN. *Knowledge-based zooming for metabolic models*, in "JOBIM", Toulouse, France, July 2013, http://hal.inria.fr/hal-00859437

Conferences without Proceedings

[23] A. ZHUKOVA, D. J. SHERMAN. *Knowledge-based generalization of metabolic networks: An applicational study*, in "Moscow Conference on Computational Molecular Biology", Moscow, Russian Federation, July 2013, http://hal.inria.fr/hal-00859440

[24] A. ZHUKOVA, D. J. SHERMAN. *What is the optimal representation of a generalized metabolic model using SBML and SBGN?*, in "COMBINE 2013", Paris, France, September 2013, http://hal.inria.fr/hal-00867373

Scientific Books (or Scientific Book chapters)

[25] P. DURRENS. *Phylogénie moléculaire des champignons*, in "Mycologie médicale", C. RIPERT (editor), Tech & Doc, Lavoisier, 2013, pp. 49-54, http://hal.inria.fr/hal-00833960

Other Publications

[26] A. ZHUKOVA. *Metabolic Model Generalization*, in "International Course in Yeast Systems Biology", Gothenburg, Sweden, June 2013, International Course in Yeast Systems Biology, http://hal.inria.fr/hal-00859442

References in notes

[27] R. ASSAR, A. GARCIA, D. J. SHERMAN. *Modeling Stochastic Switched Systems with BioRica*, in "Journées Ouvertes en Biologie, Informatique et Mathématiques JOBIM 2011", Paris, France, Institut Pasteur, July 2011, pp. 297–304, http://hal.inria.fr/inria-00617419/en

[28] R. ASSAR, A. V. LEISEWITZ, A. GARCIA, N. C. INESTROSA, M. A. MONTECINO, D. J. SHERMAN. *Reusing and composing models of cell fate regulation of human bone precursor cells*, in "BioSystems", April 2012, vol. 108, n^{o} 1-3, pp. 63-72 [*DOI :* 10.1016/J.BIOSYSTEMS.2012.01.008], http://hal.inria.fr/hal-00681022

[29] R. ASSAR, M. A. MONTECINO, D. J. SHERMAN. *Stochastic Modeling of Complex Systems and Systems Biology: From Stochastic Transition Systems to Hybrid Systems*, in "XII Latin American Congress of Probability and Mathematical Statistics", Viña del Mar, Chile, March 2012, http://hal.inria.fr/hal-00686072

[30] R. ASSAR, F. VARGAS, D. J. SHERMAN. *Reconciling competing models: a case study of wine fermentation kinetics*, in "Algebraic and Numeric Biology 2010", Austria Hagenberg, K. HORIMOTO, M. NAKATSUI, N.

POPOV (editors), Research Institute for Symbolic Computation, Johannes Kepler University of Linz, 08 2010, pp. 68–83, http://hal.inria.fr/inria-00541215/en

[31] R. ASSAR, F. VARGAS, D. J. SHERMAN. *Reconciling competing models: a case study of wine fermentation kinetics*, in "Algebraic and Numeric Biology 2010", Hagenberg, Austria, K. HORIMOTO, M. NAKATSUI, N. POPOV (editors), Lecture Notes in Computer Science, Springer, 2012, vol. 6479, pp. 68–83 [*DOI :* 10.1007/978-3-642-28067-2_6], http://hal.inria.fr/inria-00541215

[32] A. ATHANE, E. BILHÈRE, E. BON, P. LUCAS, G. MOREL, A. LONVAUD-FUNEL, C. LE HÉNAFF-LE MARREC. *Characterization of an acquired-dps-containing gene island in the lactic acid bacterium Oenococcus oeni*, in "Journal of Applied Microbiology", 2008, Received 22 October 2007, revised 8 April 2008 & Accepted 8 May 2008 (In press), http://hal.inria.fr/inria-00340058/en/

[33] R. BARRIOT, J. POIX, A. GROPPI, A. BARRE, N. GOFFARD, D. J. SHERMAN, I. DUTOUR, A. DE DARUVAR. *New strategy for the representation and the integration of biomolecular knowledge at a cellular scale*, in "Nucleic Acids Research (NAR)", 2004, vol. 32, pp. 3581-9 [*DOI :* 10.1093/NAR/GKH681], http://hal.inria.fr/inria-00202722/en/

[34] E. BON, C. GRANVALET, F. REMIZE, D. DIMOVA, P. LUCAS, D. JACOB, A. GROPPI, S. PENAUD, C. AULARD, A. DE DARUVAR, A. LONVAUD-FUNEL, J. GUZZO. *Insights into genome plasticity of the wine-making bacterium Oenococcus oeni strain ATCC BAA-1163 by decryption of its whole genome*, in "9th Symposium on Lactic Acid Bacteria", Egmond aan Zee Netherlands, 2008, http://hal.inria.fr/inria-00340073/en/

[35] L. BOURGEADE, T. MARTIN, E. BON. *PSEUDOE: A computational method to detect Psi-genes and explore PSEUDome dynamics in wine bacteria from the Oenococcus genus*, in "JOBIM2012- 13ème Journées Ouvertes en Biologie, Informatique et Mathématiques", Rennes, France, D. T. FRANÇOIS COST (editor), SFBI, Inria, July 2012, pp. 435-436, http://hal.inria.fr/hal-00722968

[36] L. BOURGEADE, T. MARTIN, A. GOULIELMAKIS, A. LONVAUD-FUNEL, P. LUCAS, E. BON. *Cracking the Pseudome Code: Inside the "silent" Psi-genes language to reconstruct Oenococcus oeni evolutionary adaptation to wine*, in "18th CBL-Club des Bactéries Lactiques Meeting", Clermont-Ferrand, France, UR CALITYSS - VETAGROSUP (editor), May 2012, 82 p. , http://hal.inria.fr/hal-00722971

[37] M. CVIJOVIC, H. SOUEIDAN, D. J. SHERMAN, E. KLIPP, M. NIKOLSKI. *Exploratory Simulation of Cell Ageing Using Hierarchical Models*, in "19th International Conference on Genome Informatics Genome Informatics", Gold Coast, Queensland Australia, J. ARTHUR, S.-K. NG (editors), Genome Informatics, Imperial College Press, London, 2008, vol. 21, pp. 114–125, EU FP6 Yeast Systems Biology Network LSHG-CT-2005-018942, EU Marie Curie Early Stage Training (EST) Network “Systems Biology”, ANR-05-BLAN-0331-03 (GENARISE), http://hal.inria.fr/inria-00350616

[38] D. DIMOVA, E. BON, P. LUCAS, R. BEUGNOT, M. DE LEEUW, A. LONVAUD-FUNEL. *The whole genome of Oenococcus strain IOEB 8413*, in "9th Symposium on Lactic Acid Bacteria", Egmond aan Zee Netherlands, 2008, http://hal.inria.fr/inria-00340086/en/

[39] A. ENACHE-ANGOULVANT, J. GUITARD, F. GRENOUILLET, T. MARTIN, P. DURRENS, C. FAIRHEAD, C. HENNEQUIN. *Rapid Discrimination between Candida glabrata, Candida nivariensis, and Candida bracarensis by Use of a Singleplex PCR*, in "Journal of Clinical Microbiology", September 2011, vol. 49, n^{o} 9, pp. 3375-3379 [*DOI :* 10.1128/JCM.00688-11], http://hal.inria.fr/inria-00625115/en

[40] A. GARCIA, D. J. SHERMAN. *Mixed-formalism hierarchical modeling and simulation with BioRica*, in "11th International Conference on Systems Biology (ICSB 2010)", United Kingdom Edimbourg, 10 2010, Poster, http://hal.inria.fr/inria-00529669/en

[41] D. GLORIAM, S. ORCHARD, D. BERTINETTI, E. BJÖRLING, E. BONGCAM-RUDLOFF, C. BORREBAECK, J. BOURBEILLON, A. R. M. BRADBURY, A. DE DARUVAR, S. DÜBEL, R. FRANK, T. J. GIBSON, L. GOLD, N. HASLAM, F. W. HERBERG, T. HILTKER, J. D. HOHEISEL, S. KERRIEN, M. KOEGL, Z. KONTHUR, B. KORN, U. LANDEGREN, L. MONTECCHI-PALAZZI, S. PALCY, H. RODRIGUEZ, S. SCHWEINSBERG, V. SIEVERT, O. STOEVESANDT, M. J. TAUSSIG, M. UEFFING, M. UHLÉN, S. VAN DER MAAREL, C. WINGREN, P. WOOLLARD, D. J. SHERMAN, H. HERMJAKOB. *A community standard format for the representation of protein affinity reagents*, in "Mol Cell Proteomics", 01 2010, vol. 9, n^{o} 1, pp. 1-10 [*DOI :* 10.1074/MCP.M900185-MCP200], http://hal.inria.fr/inria-00544751/en

[42] N. GOLENETSKAYA, D. J. SHERMAN. *Rethinking global analyses and algorithms for comparative genomics in a functional MapReduce style*, in "Algorithmique, combinatoire du texte et applications en bio-informatique (SeqBio 2011)", Lille, France, December 2011, http://hal.inria.fr/hal-00654797/en

[43] A. GOULIELMAKIS, J. BRIDIER, A. BARRÉ, O. CLAISSE, DAVID JAMES. SHERMAN, P. DURRENS, A. LONVAUD-FUNEL, E. BON. *How does Oenococcus oeni adapt to its environment? A pangenomic oligonucleotide microarray for analysis O. oeni gene expression under wine shock*, in "OENO2011- 9th International Symposium of Oenology", Bordeaux, France, P. DARRIET, L. GENY, P. LUCAS, A. LONVAUD, G. DE REVEL, P. TEISSEDRE (editors), Dunod, Paris, April 2012, pp. 358-363, http://hal.inria.fr/hal-00646867

[44] H. HERMJAKOB, L. MONTECCHI-PALAZZI, G. BADER, J. WOJCIK, L. SALWINSKI, A. CEOL, S. MOORE, S. ORCHARD, U. SARKANS, C. VON MERING, B. ROECHERT, S. POUX, E. JUNG, H. MERSCH, P. KERSEY, M. LAPPE, Y. LI, R. ZENG, D. RANA, M. NIKOLSKI, H. HUSI, C. BRUN, K. SHANKER, S. GRANT, C. SANDER, P. BORK, W. ZHU, A. PANDEY, A. BRAZMA, B. JACQ, M. VIDAL, D. J. SHERMAN, P. LEGRAIN, G. CESARENI, I. XENARIOS, D. EISENBERG, B. STEIPE, C. HOGUE, R. APWEILER. *The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data*, in "Nat. Biotechnol.", Feb. 2004, vol. 22, n^{o} 2, pp. 177-83

[45] G. JEAN, D. J. SHERMAN, M. NIKOLSKI. *Mining the semantics of genome super-blocks to infer ancestral architectures*, in "Journal of Computational Biology", 2009, http://hal.inria.fr/inria-00414692/en/

[46] N. LOIRA, T. DULERMO, M. NIKOLSKI, J.-M. NICAUD, D. J. SHERMAN. *Genome-scale Metabolic Reconstruction of the Eukaryote Cell Factory Yarrowia Lipolytica*, in "11th International Conference on Systems Biology (ICSB 2010)", United Kingdom Edimbourg, 10 2010, Poster, http://hal.inria.fr/hal-00652922

[47] N. LOIRA, D. J. SHERMAN, P. DURRENS. *Reconstruction and Validation of the genome-scale metabolic model of Yarrowia lipolytica iNL705*, in "Journée Ouvertes Biologie Informatique Mathématiques, JOBIM 2010", France Montpellier, 09 2010, http://www.jobim2010.fr/?q=fr/node/55

[48] D. J. SHERMAN, N. GOLENETSKAYA. *Addressing scaling-out challenges for comparative genomics*, in "Moscow Conference on Computational Molecular Biology", Moscow, Russian Federation, July 2011, http://hal.inria.fr/hal-00649189/en

[49] D. J. SHERMAN, P. DURRENS, E. BEYNE, M. NIKOLSKI, J.-L. SOUCIET. *Génolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts*, in "Nucleic Acids Research (NAR)", 2004,

vol. 32, GDR CNRS 2354 “Génolevures” [*DOI :* 10.1093/NAR/GKH091], http://hal.inria.fr/inria-00407519/en/

[50] D. J. SHERMAN, P. DURRENS, F. IRAGNE, E. BEYNE, M. NIKOLSKI, J.-L. SOUCIET. *Genolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts*, in "Nucleic Acids Res", 01 2006, vol. 34 [*DOI :* 10.1093/NAR/GKJ160], http://hal.archives-ouvertes.fr/hal-00118142/en/

[51] D. J. SHERMAN, N. LOIRA, N. GOLENETSKAYA. *High-performance comparative annotation*, in "Bioinformatics after next-generation sequencing", Zvenigorod Russian Federation, V. MAKEEV, G. KUCHEROV (editors), Russian Academy of Sciences, 06 2010, http://hal.inria.fr/inria-00563533/en/

[52] H. SOUEIDAN, M. NIKOLSKI, G. SUTRE. *Qualitative Transition Systems for the Abstraction and Comparison of Transient Behavior in Parametrized Dynamic Models*, in "Computational Methods in Systems Biology (CMSB’09)", Italie Bologna, Springer Verlag, 2009, vol. 5688, pp. 313–327, http://hal.archives-ouvertes.fr/hal-00408909/en/

[53] H. SOUEIDAN, D. J. SHERMAN, M. NIKOLSKI. *BioRica: A multi model description and simulation system*, in "F0SBE", Allemagne, 2007, pp. 279-287, http://hal.archives-ouvertes.fr/hal-00306550/en/

[54] N. VYAHHI, A. GOËFFON, D. J. SHERMAN, M. NIKOLSKI. *Swarming Along the Evolutionary Branches Sheds Light on Genome Rearrangement Scenarios*, in "ACM SIGEVO Conference on Genetic and evolutionary computation", F. ROTHLAUF (editor), ACM, 2009, http://hal.inria.fr/inria-00407508/en/